Empirical risk minimization in inverse problems

Jussi Klemelä
University of Mannheim

October 14, 2008

Contents

Contents

- A glimpse at empirical risk minimization (in the setting of density estimation)
- A glimpse at statistical inverse problems
- Can statistical inverse problems be solved with empirical risk minimization?

Contents

- A glimpse at empirical risk minimization (in the setting of density estimation)
- A glimpse at statistical inverse problems
- Can statistical inverse problems be solved with empirical risk minimization?

Part I: Empirical risk minimization in the direct case

- We observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f. We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Estimator $\hat{f}$ is chosen to minimize empirical risk $\gamma_n(\hat{f})$ over $\hat{f} \in C$.

- log-likelihood empirical risk:

$$\gamma_n(\hat{f}) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}(X_i)$$

- L_2 empirical risk:

$$\gamma_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \left(-2\hat{f}(X_i) + \|\hat{f}\|_2^2\right)$$

Part I: Empirical risk minimization in the direct case

- We observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f. We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Estimator $\hat{f}$ is chosen to minimize empirical risk $\gamma_n(\hat{f})$ over $\hat{f} \in C$.

- log-likelihood empirical risk:

$$\gamma_n(\hat{f}) = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}(X_i)$$

- L_2 empirical risk:

$$\gamma_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(-2\hat{f}(X_i) + \|\hat{f}\|_2^2\right)$$

Part I: Empirical risk minimization in the direct case

- We observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f. We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Estimator $\hat{f}$ is chosen to minimize empirical risk $\gamma_n(\hat{f})$ over $\hat{f} \in C$.

- log-likelihood empirical risk:

$$\gamma_n(\hat{f}) = -\frac{1}{n}\sum_{i=1}^{n} \log \hat{f}(X_i)$$

- L_2 empirical risk:

$$\gamma_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \left(-2\hat{f}(X_i) + \|\hat{f}\|_2^2\right)$$

Part I: Empirical risk minimization in the direct case

- We observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f. We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Estimator $\hat{f}$ is chosen to minimize empirical risk $\gamma_n(\hat{f})$ over $\hat{f} \in C$.

- log-likelihood empirical risk:

$$\gamma_n(\hat{f}) = -\frac{1}{n} \sum_{i=1}^{n} \log \hat{f}(X_i)$$

- L_2 empirical risk:

$$\gamma_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(-2\hat{f}(X_i) + \|\hat{f}\|_2^2 \right)$$

Empirical risk minimization without regularization

L_2 empirical risk

- L_2 empirical risk:

$$\gamma_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left(-2\hat{f}(X_i) + \|\hat{f}\|_2^2 \right).$$

- We have

$$\begin{aligned} \|\hat{f} - f\|_2^2 - \|f\|_2^2 &= -2 \int_{\mathbf{R}^d} f \hat{f} + \|\hat{f}\|_2^2 \\ &\approx -\frac{2}{n} \sum_{i=1}^{n} \hat{f}(X_i) + \|\hat{f}\|_2^2 \\ &= \gamma_n\left(\hat{f}\right). \end{aligned}$$

L_2 empirical risk

- L_2 empirical risk:

$$\gamma_n(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(-2\hat{f}(X_i) + \|\hat{f}\|_2^2\right).$$

- We have

$$\begin{aligned}
\|\hat{f} - f\|_2^2 - \|f\|_2^2 &= -2\int_{\mathbf{R}^d} f\hat{f} + \|\hat{f}\|_2^2 \\
&\approx -\frac{2}{n}\sum_{i=1}^{n}\hat{f}(X_i) + \|\hat{f}\|_2^2 \\
&= \gamma_n\left(\hat{f}\right).
\end{aligned}$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Dense minimizer $\hat{f}$ is a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{f}) \le \inf_{g \in \mathcal{F}} \gamma_n(g) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a finite δ-net of $\mathcal{F}$: for each $f \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||f - \phi||_2 \le \delta$.
  - Define the δ-net estimator $\hat{f}$ by

$$\hat{f} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Dense minimizer $\hat{f}$ is a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{f}) \leq \text{inf}_{g \in \mathcal{F}} \gamma_n(g) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a finite δ-net of $\mathcal{F}$: for each $f \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||f - \phi||_2 \leq \delta$.
  - Define the δ-net estimator $\hat{f}$ by

$$\hat{f} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Dense minimizer $\hat{f}$ is a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{f}) \leq \inf_{g \in \mathcal{F}} \gamma_n(g) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a finite δ-net of $\mathcal{F}$: for each $f \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $\|f - \phi\|_2 \leq \delta$.
  - Define the δ-net estimator $\hat{f}$ by

$$\hat{f} = \operatorname{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Dense minimizer $\hat{f}$ is a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{f}) \leq \inf_{g \in \mathcal{F}} \gamma_n(g) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a finite δ-net of $\mathcal{F}$: for each $f \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $\|f - \phi\|_2 \leq \delta$.
  - Define the δ-net estimator $\hat{f}$ by

$$\hat{f} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Dense minimizer $\hat{f}$ is a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{f}) \leq \inf_{g \in \mathcal{F}} \gamma_n(g) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a finite δ-net of $\mathcal{F}$: for each $f \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||f - \phi||_2 \leq \delta$.
  - Define the δ-net estimator $\hat{f}$ by

$$\hat{f} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Remark: Complexity penalization

The classical Sobolev space/spline case:

- Dense minimizer minimizes

$$\gamma_n(g) \text{ over } g \in \mathcal{F} = \{g : \|g''\|_2 \leq L\}.$$

- This is equivalent to minimizing

$$\gamma_n(g) + \alpha \cdot \|g''\|_2^2 \text{ over } g.$$

Remark: Complexity penalization

The classical Sobolev space/spline case:

- Dense minimizer minimizes

$$\gamma_n(g) \text{ over } g \in \mathcal{F} = \{g : \|g''\|_2 \le L\}.$$

- This is equivalent to minimizing

$$\gamma_n(g) + \alpha \cdot \|g''\|_2^2 \text{ over } g.$$

Bounds for the mean integrated squared error

- Upper bound for the dense minimizer:
  Define the entropy integral $G(\psi) = \int_0^{\psi} \sqrt{\log_e(\#\mathcal{F}_\delta)}\, d\delta$.
  Let ψ_n be such that $\psi_n^2 \geq C'\, n^{-1/2} G(\psi_n)$, for a positive constant C'.
  For $f \in \mathcal{F}$,
  $$E \left\| \hat{f} - f \right\|_2^2 \leq C \left(\psi_n^2 + \epsilon \right),$$
  for a positive constant C.

- Upper bound for the δ-net estimator: For $f \in \mathcal{F}$,
  $$E \left\| \hat{f} - f \right\|_2^2 \leq C_1 \delta^2 + C_2 \frac{\log_e(\#\mathcal{F}_\delta) + 1}{n},$$
  where C_1, C_2 are positive constants.

Bounds for the mean integrated squared error

- Upper bound for the dense minimizer:
  Define the entropy integral $G(\psi) = \int_0^{\psi} \sqrt{\log_e(\#\mathcal{F}_\delta)}\, d\delta$.
  Let ψ_n be such that $\psi_n^2 \geq C'\, n^{-1/2} G(\psi_n)$, for a positive constant C'.
  For $f \in \mathcal{F}$,
  $$E \left\| \hat{f} - f \right\|_2^2 \leq C \left(\psi_n^2 + \epsilon \right),$$
  for a positive constant C.

- Upper bound for the δ-net estimator: For $f \in \mathcal{F}$,
  $$E \left\| \hat{f} - f \right\|_2^2 \leq C_1 \delta^2 + C_2 \frac{\log_e(\#\mathcal{F}_\delta) + 1}{n},$$
  where C_1, C_2 are positive constants.

Empirical risk minimization

One may minimize the empirical risk over

- smoothness classes,
- classes defined by structural restrictions,
- parametric, semiparametric restrictions.

Empirical risk minimization

One may minimize the empirical risk over

- smoothness classes,
- classes defined by structural restrictions,
- parametric, semiparametric restrictions.

Empirical risk minimization

One may minimize the empirical risk over

- smoothness classes,
- classes defined by structural restrictions,
- parametric, semiparametric restrictions.

Part II: Statistical inverse problems: density estimation

- Direct case: we observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f. We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Inverse case: we observe $Y_1, \ldots, Y_n \in \mathbf{Y}$ which are i.i.d. with density Ag. We want to estimate function $g : \mathbf{R}^d \to \mathbf{R}$.

Part II: Statistical inverse problems: density estimation

- Direct case: we observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f. We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Inverse case: we observe $Y_1, \ldots, Y_n \in \mathbf{Y}$ which are i.i.d. with density Ag. We want to estimate function $g : \mathbf{R}^d \to \mathbf{R}$.

Example I: Deconvolution

- Let A be the convolution operator:

$$(Ag)(x) = (K * g)(x) = \int K(x-y)g(y)\,dy$$

for some function K.

- If $Y_1, \ldots, Y_n \sim Ag$, then we may write

$$Y_i = X_i + \epsilon_i,$$

where $X_i \sim g$, $\epsilon_i \sim K$, and $\epsilon_i \perp\!\!\!\perp X_i$.
The observations X_i are contaminated with measurements errors.

Example I: Deconvolution

- Let A be the convolution operator:

$$(Ag)(x) = (K * g)(x) = \int K(x-y)g(y)\,dy$$

  for some function K.

- If $Y_1, \ldots, Y_n \sim Ag$, then we may write

$$Y_i = X_i + \epsilon_i,$$

  where $X_i \sim g$, $\epsilon_i \sim K$, and $\epsilon_i \perp\!\!\!\perp X_i$.
  The observations X_i are contaminated with measurements errors.

Example II: Tomography

Example II: Tomography

The sample space of lines is identified with $\mathbf{Y} = S_d \times [0, \infty)$, since lines may be written as $P_{s,u} = \{z \in \mathbf{R}^d : z^T s = u\}$, where $s \in S_d = \{x \in \mathbf{R}^d : \|s\| = 1\}$, $u \geq 0$.

Example II: Tomography

- Let us define $Ag : S_d \times [0, \infty) \to \mathbf{R}$, where $g : \mathbf{R}^d \to \mathbf{R}$ is the underlying density.

- Let $(S, U) \sim Ag$. The basic observation is that

$$f_{U \mid S=s}(u) = \int_{P_{s,u}} g,$$

where $P_{s,u} = \{z \in \mathbf{R}^d : z^T s = u\}$, $s \in S_d$, $u \geq 0$.
In addition, $S \sim \text{Unif}(S_d)$.

- Now $(Ag)(s, u) = f_S(s) \cdot f_{U \mid S=s}(u) = \frac{1}{\mu(S_d)} \int_{P_{s,u}} g$.

- The Radon transform of $g : \mathbf{R}^d \to \mathbf{R}$ is $(Rg)(s, u) = \int_{P_{s,u}} g$.

Example II: Tomography

- Let us define $Ag : S_d \times [0, \infty) \to \mathbf{R}$,
  where $g : \mathbf{R}^d \to \mathbf{R}$ is the underlying density.

- Let $(S, U) \sim Ag$. The basic observation is that

$$f_{U \mid S=s}(u) = \int_{P_{s,u}} g,$$

  where $P_{s,u} = \{z \in \mathbf{R}^d : z^T s = u\}$, $s \in S_d$, $u \geq 0$.
  In addition, $S \sim \text{Unif}(S_d)$.

- Now $(Ag)(s, u) = f_S(s) \cdot f_{U \mid S=s}(u) = \frac{1}{\mu(S_d)} \int_{P_{s,u}} g$.

- The Radon transform of $g : \mathbf{R}^d \to \mathbf{R}$ is $(Rg)(s, u) = \int_{P_{s,u}} g$.

Example II: Tomography

- Let us define $Ag : S_d \times [0, \infty) \to \mathbf{R}$,
  where $g : \mathbf{R}^d \to \mathbf{R}$ is the underlying density.

- Let $(S, U) \sim Ag$. The basic observation is that

$$f_{U \mid S=s}(u) = \int_{P_{s,u}} g,$$

  where $P_{s,u} = \{z \in \mathbf{R}^d : z^T s = u\}$, $s \in S_d$, $u \geq 0$.
  In addition, $S \sim \text{Unif}(S_d)$.

- Now $(Ag)(s, u) = f_S(s) \cdot f_{U \mid S=s}(u) = \frac{1}{\mu(S_d)} \int_{P_{s,u}} g$.

- The Radon transform of $g : \mathbf{R}^d \to \mathbf{R}$ is $(Rg)(s, u) = \int_{P_{s,u}} g$.

Example II: Tomography

- Let us define $Ag : S_d \times [0, \infty) \to \mathbf{R}$,
  where $g : \mathbf{R}^d \to \mathbf{R}$ is the underlying density.

- Let $(S, U) \sim Ag$. The basic observation is that

$$f_{U \mid S=s}(u) = \int_{P_{s,u}} g,$$

  where $P_{s,u} = \{z \in \mathbf{R}^d : z^T s = u\}$, $s \in S_d$, $u \geq 0$.
  In addition, $S \sim \text{Unif}(S_d)$.

- Now $(Ag)(s, u) = f_S(s) \cdot f_{U \mid S=s}(u) = \frac{1}{\mu(S_d)} \int_{P_{s,u}} g$.

- The Radon transform of $g : \mathbf{R}^d \to \mathbf{R}$ is $(Rg)(s, u) = \int_{P_{s,u}} g$.

Example III: Estimation of the derivative

Let

$$(Ag)(t) = \int_{-\infty}^{t} g(u)\,du.$$

Let $X_1, \ldots, X_n$ be i.i.d. with density $f = Ag$. We have

$$g(t) = f'(t).$$

Inverse problems of applied mathematics

- We want to find $g : \mathbf{R}^d \to \mathbf{R}$ which satisfies

$$Ag = y$$

  for a given y.

- Sampling operator: $Ag = (g(x_1), \ldots, g(x_n))$. Now g is not uniquely determined.

- Convolution operator: $Ag = K * g = y$.
  - If $F(K)(\omega) > 0$ for all ω, then g is uniquely determined: We have $F(K * g) = F(K) \cdot F(g)$ and thus $g = F^{-1}(F(y)/F(K))$.

Inverse problems of applied mathematics

- We want to find $g : \mathbf{R}^d \to \mathbf{R}$ which satisfies

$$Ag = y$$

  for a given y.

- Sampling operator: $Ag = (g(x_1), \ldots, g(x_n))$. Now g is not uniquely determined.

- Convolution operator: $Ag = K * g = y$.
  - If $F(K)(\omega) > 0$ for all ω, then g is uniquely determined: We have $F(K * g) = F(K) \cdot F(g)$ and thus $g = F^{-1}(F(y)/F(K))$.

Inverse problems of applied mathematics

- We want to find $g : \mathbf{R}^d \to \mathbf{R}$ which satisfies

$$Ag = y$$

  for a given y.

- Sampling operator: $Ag = (g(x_1), \ldots, g(x_n))$. Now g is not uniquely determined.

- Convolution operator: $Ag = K * g = y$.
  - If $F(K)(\omega) > 0$ for all ω, then g is uniquely determined: We have $F(K * g) = F(K) \cdot F(g)$ and thus $g = F^{-1}(F(y)/F(K))$.

Inverse problems of applied mathematics

- We want to find $g : \mathbf{R}^d \to \mathbf{R}$ which satisfies

$$Ag = y$$

  for a given y.

- Sampling operator: $Ag = (g(x_1), \ldots, g(x_n))$. Now g is not uniquely determined.

- Convolution operator: $Ag = K * g = y$.

  - If $F(K)(\omega) > 0$ for all ω, then g is uniquely determined: We have $F(K * g) = F(K) \cdot F(g)$ and thus $g = F^{-1}(F(y)/F(K))$.

Well posed inverse problems

An inverse problem is well-posed, when

1. the solution exists
2. it is unique
3. it is stable: it depends continuously on data:
   A^{-1} has to be continuous.

Part III: Can inverse problems be solved with empirical risk minimization?

- Direct case: we observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f.
  We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Inverse case: we observe $Y_1, \ldots, Y_n \in \mathbf{Y}$ which are i.i.d. with density Ag.
  We want to estimate function $g : \mathbf{R}^d \to \mathbf{R}$.

Part III: Can inverse problems be solved with empirical risk minimization?

- Direct case: we observe $X_1, \ldots, X_n \in \mathbf{R}^d$ which are i.i.d. with density f. We want to estimate $f : \mathbf{R}^d \to \mathbf{R}$.

- Inverse case: we observe $Y_1, \ldots, Y_n \in \mathbf{Y}$ which are i.i.d. with density Ag. We want to estimate function $g : \mathbf{R}^d \to \mathbf{R}$.

A classical solution

Find $\hat{g}$ minimizing

$$\gamma_n(A\hat{g}) + \alpha \cdot D(\hat{g}),$$

where

$$\gamma_n(A\hat{g}) = -\frac{1}{n}\sum_{i=1}^{n} \log(A\hat{g})(Y_i)$$

and $D(g)$ is a penalization.
For example, $D(g) = \|g''\|_2^2$.

A new solution

Define the L_2 empirical risk by

$$\gamma_n(g) = \frac{1}{n} \sum_{i=1}^{n} \left(-2(Qg)(Y_i) + \|g\|_2^2 \right),$$

where Q is the adjoint of the inverse of A:

$$\int_{\mathbf{R}^d} (A^{-1}h)g = \int_{\mathbf{Y}} h(Qg),$$

for $h \in L_2(\mathbf{Y})$, $g \in L_2(\mathbf{R}^d)$.
Remember the direct case: $\gamma_n(g) = \frac{1}{n} \sum_{i=1}^{n} \left(-2g(X_i) + \|g\|_2^2 \right)$.

L_2 empirical risk; motivation

The L_2 empirical risk:

$$\gamma_n(\hat{g}) = n^{-1} \sum_{i=1}^{n} \left(-2(Q\hat{g})(Y_i) + \|\hat{g}\|_2^2 \right).$$

Now,

$$\begin{aligned} \|\hat{g} - g\|_2^2 - \|g\|_2^2 &= -2 \int_{\mathbf{R}^d} g\hat{g} + \|\hat{g}\|_2^2 = -2 \int_{\mathbf{R}^d} (A^{-1}Ag)\hat{g} + \|\hat{g}\|_2^2 \\ &= -2 \int_{\mathbf{Y}} (Ag)(Q\hat{g}) + \|\hat{g}\|_2^2 \approx -\frac{2}{n} \sum_{i=1}^{n} (Q\hat{g})(Y_i) + \|\hat{g}\|_2^2 \\ &= \gamma_n(\hat{g}). \end{aligned}$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Define the dense minimizer $\hat{g}$ as a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{g}) \leq \inf_{h \in \mathcal{F}} \gamma_n(h) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a δ-net of $\mathcal{F}$: for each $g \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||g - \phi||_2 \leq \delta$.
  - Define the δ-net estimator $\hat{g}$ by

$$\hat{g} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Define the dense minimizer $\hat{g}$ as a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{g}) \leq \text{inf}_{h \in \mathcal{F}} \gamma_n(h) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a δ-net of $\mathcal{F}$: for each $g \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||g - \phi||_2 \leq \delta$.
  - Define the δ-net estimator $\hat{g}$ by

$$\hat{g} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Define the dense minimizer $\hat{g}$ as a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{g}) \leq \inf_{h \in \mathcal{F}} \gamma_n(h) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a δ-net of $\mathcal{F}$: for each $g \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||g - \phi||_2 \leq \delta$.
  - Define the δ-net estimator $\hat{g}$ by

$$\hat{g} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Define the dense minimizer $\hat{g}$ as a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:
$$\gamma_n(\hat{g}) \leq \text{inf}_{h\in\mathcal{F}}\gamma_n(h) + \epsilon.$$
- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a δ-net of $\mathcal{F}$: for each $g \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||g - \phi||_2 \leq \delta$.
  - Define the δ-net estimator $\hat{g}$ by
$$\hat{g} = \text{argmin}_{\phi\in\mathcal{F}_\delta}\gamma_n(\phi).$$

Dense minimizer and δ-net minimizer

- Dense minimizer
  - Define the dense minimizer $\hat{g}$ as a minimizer of the empirical risk over $\mathcal{F}$, up to $\epsilon > 0$:

$$\gamma_n(\hat{g}) \leq \text{inf}_{h \in \mathcal{F}} \gamma_n(h) + \epsilon.$$

- δ-net minimizer
  - Let $\mathcal{F}_\delta$ be a δ-net of $\mathcal{F}$: for each $g \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $||g - \phi||_2 \leq \delta$.
  - Define the δ-net estimator $\hat{g}$ by

$$\hat{g} = \text{argmin}_{\phi \in \mathcal{F}_\delta} \gamma_n(\phi).$$

Bounds for the mean integrated squared error

- Upper bound for dense minimizer: Define the operator norm

$$T_\delta = \max_{\phi,\phi' \in \mathcal{F}_\delta, \phi \neq \phi'} \frac{\|Q(\phi - \phi')\|_2}{\|\phi - \phi'\|_2}, \qquad \delta > 0.$$

  Define the entropy integral $G(\psi) = \int_0^\psi T_\delta \sqrt{\log_e(\#\mathcal{F}_\delta)}\, d\delta$.
  Let ψ_n be such that $\psi_n^2 \geq C'\, n^{-1/2} G(\psi_n)$, for a positive constant C'.
  For $g \in \mathcal{F}$,

$$E\,\|\hat{g} - g\|_2^2 \leq C\left(\psi_n^2 + \epsilon\right),$$

  for a positive constant C.

- Upper bound for δ-net estimator: For $g \in \mathcal{F}$,

$$E\,\|\hat{g} - g\|_2^2 \leq C_1 \delta^2 + C_2 \frac{T_\delta^2 \cdot (\log_e(\#\mathcal{F}_\delta) + 1)}{n}.$$

Bounds for the mean integrated squared error

- Upper bound for dense minimizer: Define the operator norm

$$T_\delta = \max_{\phi,\phi' \in \mathcal{F}_\delta, \phi \neq \phi'} \frac{\|Q(\phi - \phi')\|_2}{\|\phi - \phi'\|_2}, \qquad \delta > 0.$$

  Define the entropy integral $G(\psi) = \int_0^\psi T_\delta \sqrt{\log_e(\#\mathcal{F}_\delta)}\, d\delta$.
  Let ψ_n be such that $\psi_n^2 \geq C' \, n^{-1/2} G(\psi_n)$, for a positive constant C'.
  For $g \in \mathcal{F}$,

$$E\, \|\hat{g} - g\|_2^2 \leq C \left(\psi_n^2 + \epsilon\right),$$

  for a positive constant C.

- Upper bound for δ-net estimator: For $g \in \mathcal{F}$,

$$E\, \|\hat{g} - g\|_2^2 \leq C_1 \delta^2 + C_2 \frac{T_\delta^2 \cdot (\log_e(\#\mathcal{F}_\delta) + 1)}{n}.$$

Universal rate formula

- The minimax rate ψ_n over $\mathcal{F}$ is the solution to the equation

$$n\psi_n^2 = T_{\psi_n}^2 \log \#\mathcal{F}_{\psi_n},$$

where

– $\mathcal{F}_\delta$ is a minimal δ-net of $\mathcal{F}$
(for each $f \in \mathcal{F}$ there is $\phi \in \mathcal{F}_\delta$ such that $\|f - \phi\|_2 \leq \delta$)

– and T_δ is the operator norm

$$T_\delta = \max_{\phi,\phi' \in \mathcal{F}_\delta, \phi \neq \phi'} \frac{\|Q(\phi - \phi')\|_2}{\|\phi - \phi'\|_2}, \qquad \delta > 0.$$

Examples of MISE bounds

- Let $\log(\#\mathcal{F}_\delta) \asymp \delta^{-b}$ and $T_\delta \asymp \delta^{-a}$.
  For a smoothness class with smoothness index s we have $b = d/s$.

- Universal rate formula:

$$n\psi_n^2 = T_{\psi_n}^2 \log(\#\mathcal{F}_{\psi_n}) \Leftrightarrow \psi_n = n^{-1/[2(a+1)+b]}.$$

- For the dense minimizer: $G(\psi) \asymp \int_0^\psi \delta^{-a-b/2}\, d\delta \asymp \psi^{-a-b/2+1}$ when $a + b/2 < 1$.
  The rate is given by

$$\psi_n^2 = n^{-1/2} G(\psi_n) \Leftrightarrow \psi_n = n^{-1/[2(a+1)+b]}.$$

Examples of MISE bounds

- Let $\log(\#\mathcal{F}_\delta) \asymp \delta^{-b}$ and $T_\delta \asymp \delta^{-a}$.
  For a smoothness class with smoothness index s we have $b = d/s$.

- Universal rate formula:

$$n\psi_n^2 = T_{\psi_n}^2 \log(\#\mathcal{F}_{\psi_n}) \Leftrightarrow \psi_n = n^{-1/[2(a+1)+b]}.$$

- For the dense minimizer: $G(\psi) \asymp \int_0^\psi \delta^{-a-b/2}\, d\delta \asymp \psi^{-a-b/2+1}$ when $a + b/2 < 1$.
  The rate is given by

$$\psi_n^2 = n^{-1/2} G(\psi_n) \Leftrightarrow \psi_n = n^{-1/[2(a+1)+b]}.$$

Examples of MISE bounds

- Let $\log(\#\mathcal{F}_\delta) \asymp \delta^{-b}$ and $T_\delta \asymp \delta^{-a}$.
  For a smoothness class with smoothness index s we have $b = d/s$.

- Universal rate formula:

$$n\psi_n^2 = T_{\psi_n}^2 \log(\#\mathcal{F}_{\psi_n}) \Leftrightarrow \psi_n = n^{-1/[2(a+1)+b]}.$$

- For the dense minimizer: $G(\psi) \asymp \int_0^\psi \delta^{-a-b/2}\, d\delta \asymp \psi^{-a-b/2+1}$ when $a + b/2 < 1$.
  The rate is given by

$$\psi_n^2 = n^{-1/2} G(\psi_n) \Leftrightarrow \psi_n = n^{-1/[2(a+1)+b]}.$$

Examples of Q (adjoint of the inverse of A)

- Deconvolution: Let $Ag = K * g$. Then $Qg = F^{-1}(Fg/FK)$.
- Radon transform: Let $(Ag)(s,u) = \int_{P_{s,u}} g$. Then

$$(Qg)(s,u) = C \cdot (F_1^{-1} I_s g)(u),$$

where $I_s g(t) = (Fg)(ts)$, for $s \in S_d$, $t \in [0, \infty)$.

Examples of Q (adjoint of the inverse of A)

- Deconvolution: Let $Ag = K * g$. Then $Qg = F^{-1}(Fg/FK)$.

- Radon transform: Let $(Ag)(s,u) = \int_{P_{s,u}} g$. Then

$$(Qg)(s,u) = C \cdot (F_1^{-1} I_s g)(u),$$

where $I_s g(t) = (Fg)(ts)$, for $s \in S_d$, $t \in [0,\infty)$.

An alternative I: Singular value decomposition

Let Φ be the basis consisting of eigenfunctions of operator A^*A (A^* is the adjoint operator to the operator A).
The singular value decomposition estimator is defined by

$$\hat{g}(x) = \sum_{\phi \in \Phi_n} \gamma_\phi^{-1} \hat{\psi}_{\phi,n} \phi(x),$$

where $\Phi_n \subset \Phi$ is finite,
γ_ϕ^2 is the eigenvalue corresponding to ϕ: $A^*A\phi = \gamma_\phi^2 \phi$,

$$\hat{\psi}_{\phi,n} \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} \psi_\phi(Y_i) \approx < Ag, \psi_\phi >,$$

$$\psi_\phi = A\phi / \|A\phi\|_2.$$

Motivation of the singular value decomposition estimator

Let Φ be the basis consisting of eigenfunctions of operator A^*A. Denote with γ_ϕ^2 the eigenvalue corresponding to ϕ: $A^*A\phi = \gamma_\phi^2\phi, \qquad (A^*A)^{-1}\phi = \gamma_\phi^{-2}\phi$.
Denote $\psi_\phi = A\phi/\|A\phi\|_2$. We may write

$$
\begin{aligned}
g &= \sum_{\phi\in\Phi} < g,\phi > \phi = \sum_{\phi\in\Phi} < (A^*A)^{-1}A^*Ag,\phi > \phi \\
&= \sum_{\phi\in\Phi} \gamma_\phi^{-2} < A^*Ag,\phi > \phi = \sum_{\phi\in\Phi} \gamma_\phi^{-2} < Ag, A\phi > \phi \\
&= \sum_{\phi\in\Phi} \gamma_\phi^{-1} < Ag,\psi_\phi > \phi,
\end{aligned}
$$

where we used the fact

$$
\|A\phi\|_2^2 =< A\phi, A\phi >=< \phi, A^*A\phi >= \gamma_\phi^2\|\phi\|_2^2 = \gamma_\phi^2.
$$

An alternative II: Wavelet-vaguelette decomposition

The wavelet-vaguelette estimator is defined by

$$\hat{g}(x) = \sum_{\phi \in \Phi_n} \beta_\phi^{-1} \hat{u}_{\phi,n} \phi(x),$$

where $\Phi_n \subset \Phi$,
Φ is a wavelet basis,

$$\hat{u}_{\phi,n} \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} u_\phi(Y_i) \approx < Ag, u_\phi >,$$

u_ϕ satisfies

$$A^* u_\phi = \beta_\phi \phi.$$

Motivation of the wavelet-vaguelette estimator

Let Φ be a wavelet basis. Let u_ϕ satisfy

$$A^* u_\phi = \beta_\phi \phi.$$

We may write

$$\begin{aligned} g &= \sum_{\phi \in \Phi} < g, \phi > \phi \\ &= \sum_{\phi \in \Phi} < g, A^* u_\phi > \beta_\phi^{-1} \phi \\ &= \sum_{\phi \in \Phi} < Ag, u_\phi > \beta_\phi^{-1} \phi. \end{aligned}$$

Summary

- L_2 empirical risk may be used to solve statistical inverse problems
- estimators are adapted to the underlying function
- a comprehensive mathematical analysis of the behavior of the estimators is possible

Summary

- L_2 empirical risk may be used to solve statistical inverse problems
- estimators are adapted to the underlying function
- a comprehensive mathematical analysis of the behavior of the estimators is possible

Summary

- L_2 empirical risk may be used to solve statistical inverse problems
- estimators are adapted to the underlying function
- a comprehensive mathematical analysis of the behavior of the estimators is possible