Less Exact Models

can yield

More Exact Solutions

Howard Heaton

Typal Academy

Model Approximation

A constrained optimization problem may be written as

$$\min_x f(x) \quad \text{s.t.} \quad x \in \mathcal{C}, \tag{P}$$

where f is the objective and $\mathcal{C}$ is the constraint. For various reasons, it is common to approximate (P) by

$$\min_x f_\varepsilon(x) \quad \text{s.t.} \quad x \in \mathcal{C}_\varepsilon, \tag{P$_\varepsilon$}$$

where f_ε is an approximation of f and $\mathcal{C}_\varepsilon$ an approximation of $\mathcal{C}$. Ideally, the solution $x^\star$ to (P) is well-approximated by the solution $x_\varepsilon^\star$ to (P$_\varepsilon$).

Algorithm Approximation

Typically, solutions to large optimization problems are estimated numerically via iterative procedures. For example, if f is differentiable, projected gradient constructs a sequence $\{x^k\}$ of solution estimates via

$$x^{k+1} = \text{proj}_{\mathcal{C}}\left(x - \alpha \nabla f(x)\right),$$

where $\alpha > 0$ is a step size and $\text{proj}_{\mathcal{C}}$ is the Euclidean projection onto $\mathcal{C}$. With suitable assumptions,

$$\lim_{k \to \infty} x^k = x^\star.$$

In practice, a finite index K is chosen with $x^K \approx x^\star$.

Solution Estimate Error

If one uses an approximate model and an approximate algorithm, then there are two sources of error. That is, the output x_ε^K of an algorithm for solving (P_ε) has

$$\begin{aligned}(\text{estimate error}) &= x_\varepsilon^K - x^\star \\ &= (x_\varepsilon^K - x_\varepsilon^\star) + (x_\varepsilon^\star - x^\star) \\ &= (\text{algorithm error}) + (\text{model error}).\end{aligned}$$

When $x^\star$ is not known, one may consider other factors:

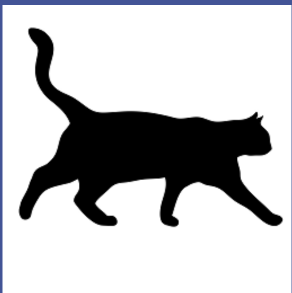
$$(\text{constraint violation}) = \mathrm{dist}(x_\varepsilon^K, \mathcal{C}) = \min_{z\in\mathcal{C}} \|z - x_\varepsilon^K\|$$

or

$$(\text{objective suboptimality}) = f(x_\varepsilon^K) - f(x^\star).$$

Example - Earth Mover's Distance

The earth mover's distance (EMD) is a key metric that is widely used in several fields. It measures the distance between a distribution ρ^0 and ρ^1. In this example, we let ρ^0 and ρ^1 be cat images.



ρ^0 = Standing Cat

ρ^1 = Crouching Cat

Example - EMD Formulation

The EMD[†] can be characterized as the optimal objective value for the problem

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \underbrace{\text{div}(x) + \rho^1 - \rho^0 = 0,}_{\mathcal{C}} \qquad \text{(P)}$$

where div denotes a linear operation (think "matrix").

Picking $\varepsilon = 10^{-10}$, an approximate version is

$$\min_x \|x\|_1 \quad \text{s.t.} \quad \underbrace{\|\text{div}(x) + \rho^1 - \rho^0 = 0\| \leq \varepsilon.}_{\mathcal{C}_\varepsilon} \qquad (\text{P}_\varepsilon)$$

The inequality constraint in (P_ε) changes the structure of the problem and, thus, what algorithms can be used.

[†] Specifcally, we use the Wasserstein-1 distance here.

Example - EMD Algorithms

Primal-dual hybrid gradient (PDHG) solves (P):

- first-order method with efficient updates
- converges to optimal solution
- estimates satisfy constraint asymptotically $\mathcal{C}$

Here "asymptotically" means $\lim_{k\to\infty} \text{dist}(x^k, \mathcal{C}) = 0$.

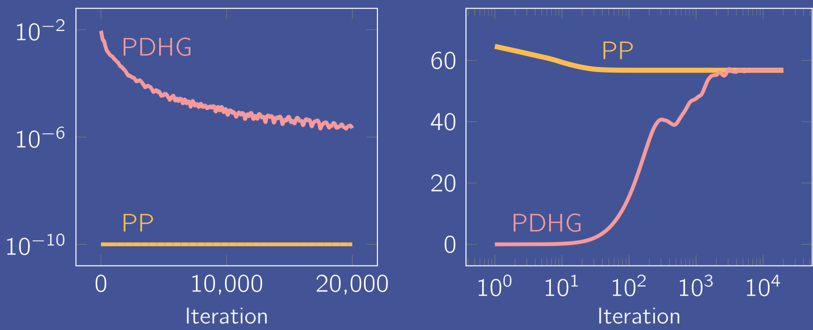
Proximal projection (PP) algorithm solves (P_ε):

- first-order method with efficient updates
- converges to optimal solution
- each estimate x_ε^k satisfies constraint $\mathcal{C}_\varepsilon$

Note: PP only works in this setting when $\varepsilon > 0$

Example - Convergence Plots

Violation $\|\text{div}(x^k) + \rho^1 - \rho^0\|_F$ **Objective** $\|x^k\|_1$



Observations:

- PP takes ~2.5X as long per step as PDHG
- PDHG requires orders of magnitude more steps
- Violation with PP is orders of magnitude lower

Takeaway: PP generates a better estimate of $x^\star$ than PDHG even though PP solves (P_ε) rather than (P)

When Inexact can be Better

In the example, the updates for PP are only defined when $\varepsilon > 0$. Thus, picking small ε "unlocked" the ability to use PP for estimating EMDs.

More generally, inexact (P_ε) may be better to use when

- (P_ε) has "nicer" structure
- (P_ε) enables circumvention of ill-conditioning and errors due to floating point arithmetic
- (P_ε) has parameter $\varepsilon > 0$ that for which (P_ε) becomes (P) as $\varepsilon \to 0^+$

Reference: *Proximal Projection Method for Stable Linearly Constrained Optimization*

Found this useful?

- ➕ Follow for more
- ♻️ Repost to share with friends